

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

APPLICATION FOR LETTERS PATENT

**Dynamic Monitor and Controller of  
Availability of a Load-Balancing Cluster**

Inventor(s):

Christopher Darling

Michael Gernaey

Hallvard Kaldestad

Howard Aikins

ATTORNEY'S DOCKET NO. MS1-681US

1  
2 **TECHNICAL FIELD**

3 This invention generally relates to a technology for remotely and  
4 dynamically monitoring the availability of the members of a load-balancing  
5 cluster. This invention further generally relates to a technology for remotely and  
6 dynamically controlling the availability of the members of a load-balancing  
7 cluster.

8  
9 **BACKGROUND**

10 During the early days of the blossoming of the World Wide Web  
11 manifestation of the Internet, there was a one-to-one relationship between Web  
12 site and computer. For each Web site, there was a single computer (generally  
13 called a "Web server") that hosted the Web site. The Web site had a single  
14 address (called an IP address) and that address was associated with the site's  
15 single computer.

16 The popularity of the Internet has become ubiquitous. Web sites are big  
17 business. For many Web sites, a single computer does not serve the volumes of  
18 activity that currently takes place and certainly cannot scale to handle the volumes  
19 to come. To accommodate these needs, the concept of "load-balancing clusters"  
20 was introduced.

21  
22 **Clustering**

23 Clustering is the connecting of two or more computers together in such a  
24 way that they behave like a single computer. Clustering is used for parallel  
25 processing, for load balancing, and for fault tolerance.

1 In general, the goal of a cluster is to make it possible to share a computing  
2 load over several systems without either the users or system administrators  
3 needing to know that more than one system is involved. If any component in the  
4 system, hardware or software fails, the user may see degraded performance, but  
5 will not lose access to the service. Ideally, if more processing power is needed, the  
6 system administrator simply "plugs in a new component", and presto, the  
7 performance of the clustered system as a whole improves.

### 8 **Load Balancing**

9  
10 As the name implies, "load balancing" attempts to evenly distribute  
11 workload amongst several computers. For Web sites, load balancing helps solve  
12 the "server-busy" problem that arises when servers drown from a flash flood of  
13 users hitting them. Load balancing prevents the problem by keeping track of  
14 which server in a group of servers user requests have been routed to and knowing  
15 roughly how busy each server is. By approximating that, load balancing  
16 determines where to direct the next request.

17 For example, a company can devote a Web site to the sporting event that it  
18 sponsors and use load balancing to handle the crush of hits during the event.  
19 Companies find load balancing useful because it is an intelligent and affordable  
20 method for apportioning high volumes of requests for server access across  
21 multiple machines, be they on the Web or in a data center.

22 With this technology, server failures are simpler to mask, reducing  
23 downtime for the end-user. Previously, managing server failures meant taking  
24 hosts out of DNS (Domain Name System) or rebooting them immediately. With  
25

1 load balancing, the failing server can be left in the failed mode for debugging  
2 without impacting end-user availability.

### 3 4 **Conventional Load-Balancing Clusters**

5 Fig. 1 illustrates a conventional load-balancing cluster 100, which consists  
6 of cluster nodes 112a-f. Typically, node 112a-f are nearly identical. Members of a  
7 cluster are referred to as nodes or servers. These terms are used interchangeably,  
8 herein.

9 In this conventional load-balancing cluster 100, a node manager 110 serves  
10 as the gatekeeper and the proxy for the nodes of the cluster. In the case of a Web  
11 site, the node manager 110 hosts the single IP address for the Web site, but it  
12 directs users to any one of the nodes 112a-f for service. Other conventional load-  
13 balancing clusters employ a partially or fully distributed scheme for managing  
14 load-balancing. An example of a fully distributed architecture is the Microsoft®  
15 Network Load-Balancing (NLB) cluster architecture. Those of ordinary skill in the  
16 art understand the existing architectures of load-balancing schemes.

17 Typically, the load-balancing cluster 100 balances the load of TCP or UDP  
18 traffic. However, end-users do not care about availability at the protocol layer.  
19 Rather, end-users care about application-layer availability. A user (such as one on  
20 clients 132-138) sends a request for information at the IP address of the cluster  
21 100 via the Internet 120. For an NLB cluster, all hosts receive this request and  
22 based on previous "convergence" criteria of the cluster, one host responds. Load is  
23 balanced statistically by subdividing the IP:port space of clients among the nodes.  
24 In the aggregate, this achieves a balanced load.

1 Load-balancing clusters provide seamless fault-tolerance in the case of  
2 server or network failures. Load-balancing cluster nodes have several specific  
3 cluster-states. For example, in NLB, those states are:

- 4 • *Suspended* – the node is not active in the cluster. It cannot be made  
5 active without an explicit “resume” request. Resume places the node in  
6 the *Stopped* state.
- 7 • *Stopped* – the node is not active in the cluster.
- 8 • *Converging* – the node is currently becoming active in the cluster. More  
9 precisely, all nodes (even those already active) move to this state any  
10 time the membership of the cluster changes.
- 11 • *Draining* – the node is not receiving new load (e.g., user requests), but  
12 existing connections are allowed to complete.
- 13 • *Converged* – the node is active in the cluster.

14  
15 More generally, load-balancing cluster nodes have these activity-related  
16 cluster-states:

- 17 • *Active* – the node is active when it is fully participating member of the  
18 cluster upon restart of the node. For example in NLB, the desired state  
19 upon restart of the node is “converged.”
  - 20 • *Inactive* – the node is inactive when it is not a participating member of  
21 the cluster upon restart of the node. For example in NLB, the desired  
22 state upon restart of the node is “Stopped.” Other examples of the  
23 inactive state include when a node is stopped or draining.
- 24  
25

Those of ordinary skill in the art understand and appreciate the conventional structure and function of a load-balancing cluster like that illustrated in Fig. 1.

### **Local and Remote Application-Layer Availability Monitoring**

Application-layer refers to the well-known OSI model. Since the application layer is the top layer, any delays at the lower layers ripple up to the application level. In addition, any errors at the lower levels impact the application layer adversely. Thus, monitoring at the application layer gives the true picture of node availability.

Herein, the focus is upon application-layer monitoring as opposed to other kinds of monitoring. An example of application-layer monitoring is performing an http GET for a Web server. An example of another type of monitoring include: checking whether Microsoft® Internet Information Server (IIS) is running as a service under Microsoft® Windows NT®; and collecting performance monitor (perfmon) counters for IIS. To the end-user, application-layer monitoring is superior for determining the actual availability of the service to an end-user.

There are two main ways to monitor application-layer availability of the nodes in a cluster: locally and remotely. Local application-layer monitoring is done from within the cluster. It is performed by the node manager and/or the nodes themselves. For example, if node manager 110 monitored the availability of the nodes 112a-f, then this is local monitoring. This type of monitoring may be called “endocluster” application-layer monitoring.

Remote application-layer monitoring is done from outside the cluster. It is not performed by the node manager and/or the nodes themselves. Rather, it is

1 performed by a computer outside of the cluster, but coupled to the cluster via a  
2 network connection. For example, if client 132 monitored the availability of the  
3 nodes 112a-f, then this is remote monitoring. This type of monitoring may be  
4 called “exocluster” application-layer monitoring. Exocluster application-layer  
5 monitoring provides a more accurate measurement of the actual availability of the  
6 nodes in the cluster than local monitoring. Why? The ultimate measure of the  
7 availability of a node is how it appears to a client from outside the cluster, such as  
8 client 132. Therefore, exocluster application-layer monitoring is better because it  
9 views node availability from the client’s perspective. Herein, this form of  
10 monitoring may also be called “client-perspective” application-layer monitoring.

11 Local application-layer monitoring is not sufficient because the systems are  
12 monitoring themselves from their own point of view. The monitoring does not  
13 follow the full path through all of the layers of the OSI model to get to the top  
14 layer—the application layer. Herein, this form of monitoring (i.e., local  
15 application-level) may also be called “cluster-perspective” application-layer  
16 monitoring.

17 Those of ordinary skill in the art are familiar with local and remote  
18 monitoring of node availability at the application-layer and are familiar with the  
19 advantages and disadvantages of both. Examples of conventional remote  
20 application-layer monitoring products include SiteScope® by Freshwater  
21 Software®.

### 22 **Limitations of Conventional Exocluster Application-Layer Monitors**

24 Passive Monitors. Conventional exocluster application-layer monitors are  
25 purely passive monitors. They are unable to actively control the nodes that they

1 are monitoring. They cannot stop a problem node. Moreover, they are unable to  
2 start an inactive node once its problems have been resolved.

3 Protocol Specific. Conventional exocluster application-layer monitors are  
4 protocol specific. They monitor defined protocols (such as HTTP and SMTP) and  
5 are incapable of monitoring other protocols without being reprogrammed.

6 Static Cluster Membership. Conventional exocluster application-layer  
7 monitors monitor a static set of hosts; there is no notion of a cluster. That is, they  
8 are not cluster-aware. They are not dynamic. In other words, they cannot  
9 dynamically monitor all of the members of the cluster as members are added and  
10 removed. They can monitor new members (or stop monitoring old members) once  
11 the membership is *statically* defined specifically for the monitor. However, the  
12 conventional exocluster application-layer monitors cannot dynamically begin  
13 monitoring new members as they are added to the cluster or dynamically stop  
14 monitoring old members as they are removed.

## 15 SUMMARY

16  
17 Described herein is a technology for remotely and dynamically monitoring  
18 the availability of the members of a load-balancing cluster. The technology  
19 provides a dynamic, exocluster application-layer monitor for dynamically  
20 monitoring and/or dynamically controlling the members of a load-balancing  
21 cluster.

22 The exocluster application-layer monitor is an active monitor—a controller.  
23 It may actively control the members that it monitors. The exocluster application-  
24 layer monitor is protocol agnostic. The exocluster application-layer monitor can  
25



1 dynamically adjust so that it can monitor all of the members of the cluster as  
2 members are added and removed.

3 This summary itself is not intended to limit the scope of this patent.  
4 Moreover, the title of this patent is not intended to limit the scope of this patent.  
5 For a better understanding of the present invention, please see the following  
6 detailed description and appending claims, taken in conjunction with the  
7 accompanying drawings. The scope of the present invention is pointed out in the  
8 appending claims.

### 9 **BRIEF DESCRIPTION OF THE DRAWINGS**

10  
11 The same numbers are used throughout the drawings to reference like  
12 elements and features.

13 Fig. 1 is a schematic diagram showing a network environment (including a  
14 cluster) within which an implementation of the invention claimed herein may be  
15 so implemented.

16 Fig. 2 is a schematic block diagram showing an embodiment of the  
17 invention claimed herein.

18 Fig. 3 is a flow diagram showing a methodological implementation of the  
19 invention claimed herein.

20 Fig. 4 is an example of a computing operating environment capable of  
21 implementing an implementation (wholly or partially) of the invention claimed  
22 herein.

## **DETAILED DESCRIPTION**

The following description sets forth specific embodiments of a dynamic monitor and controller of availability of a load-balancing cluster that incorporates elements recited in the appended claims. These embodiments are described with specificity in order to meet statutory written description, enablement, and best-mode requirements. However, the description itself is not intended to limit the scope of this patent.

Described herein are one or more exemplary implementations of a dynamic monitor and controller of availability of a load-balancing cluster. The inventors intend these exemplary implementations to be examples. The inventors do not intend these exemplary implementations to limit the scope of the claimed present invention. Rather, the inventors have contemplated that the claimed present invention might also be embodied and implemented in other ways, in conjunction with other present or future technologies.

An example of an embodiment of a dynamic monitor and controller of availability of a load-balancing cluster may be referred to as an “exemplary monitor/controller.”

## **Introduction**

The one or more exemplary implementations, described herein, of the present claimed invention may be implemented (in whole or in part) by an exocluster application-layer monitoring/controlling system 205 and/or by a computing environment like that shown in Fig. 4. More specifically, it may be

1 implemented as a program module on a “client” computer, such as client 134 of  
2 Fig. 1.

3 Particularly when a business’ Web site is their store front, they want to  
4 keep their doors open to all customers all of the time. Therefore, the Web site of  
5 such a business must be up 24/7/365 (i.e., every minute of the year) and usage  
6 delays must be minimized. Hosting the Web site on a load-balancing cluster goes  
7 a long way to achieving the goal, but some users may experience delays when a  
8 cluster member becomes overloaded or overwhelmed. To improve the overall  
9 availability of members (i.e., nodes) of a load-balancing cluster, node availability  
10 monitoring is desirable.

11 More than that, it is desirable to have the ability to automatically take  
12 corrective actions when application-layer monitoring indicates an error condition  
13 has occurred on one or more nodes. These types of application failures can be  
14 masked through manual intervention by issuing a command to remove the failing  
15 server from cluster participation, a process which takes seconds to complete  
16 (assuming that the load-balancing cluster allows such action). However, with  
17 conventional monitoring systems, the cluster is left exposed to faults occurring at  
18 the port and application level when manual intervention isn’t forthcoming. With  
19 the exemplary monitor/controller, these fault conditions can be detected and  
20 automatically remedied, without user intervention.

21 Although the exemplary monitor/controller is remote from the load-  
22 balancing cluster (i.e., “exocluster”), it integrates with the cluster for cluster-state  
23 monitoring and control. It tracks changes in cluster membership and the cluster-  
24 state of the members. With control, the exemplary monitor/controller can use  
25 application-level monitoring results to decide when to add and remove servers

1 from active cluster participation. By committing this information to a persistent  
2 store, the information needed to track the server's availability is accumulated over  
3 time.

4 Exemplary monitor/controller is a general-purpose monitoring and control  
5 system, which tests a node (of a cluster) at the application layer in order to  
6 determine whether the node is available. Exemplary monitor/controller may  
7 support any kind of testing as long as it can be captured as an app-monitor.

8 The exemplary monitor/controller tracks historical service availability. The  
9 information tracked helps answer many questions for the system administrator.  
10 Examples of such questions include:

- 11 • How many errors have the servers had today? What kinds of errors are  
12 occurring?
- 13 • What is the application-level availability of the site over the last two  
14 weeks?
- 15 • When should the capacity of the site be increased?

### 16 Some Terminology

17  
18 *SIP* – Single IP technology. A node manager has a single IP (SIP) for the  
19 entire cluster, or the nodes of a fully distributed cluster share the SIP. Thus, it  
20 uses SIP technology. Generally used in reference to either the technology or an  
21 API to programmatically connect to the technology.

22 *Cluster* – A collection of servers collectively acting as a single unit with a  
23 Single IP technology. In other words, a group of independent computer systems,  
24 referred to as nodes, working together as a unified computing resource.  
25

1        *VIP* – Virtual IP address. The IP address assigned to the SIP technology  
2 that is put into DNS.

3        *Load balancing* – A technique for scaling the performance of a server-  
4 based application (such as a Web server) by distributing its client requests across  
5 multiple nodes within the cluster. Distributing processing and communications  
6 activity evenly across a computer network so that no single device is  
7 overwhelmed. Load balancing is especially important for networks where it's  
8 difficult to predict the number of requests that will be issued to a server. Busy  
9 Web sites typically employ two or more Web servers in a load-balancing scheme.  
10 If one server starts becomes overloaded, requests are forwarded to another server  
11 with more capacity.

12        *Query* – A command verb instructing the SIP technology to return the  
13 cluster state of one or more cluster members. No changes to state are made as a  
14 result of this command.

15        *Drain* – A command verb instructing the SIP technology to prevent new  
16 connections to the specified server or cluster, while allowing existing connections  
17 to “bleed” off. In other words, a state in which a node is no longer accepting  
18 incoming traffic is draining. No new connections are allowed, but existing  
19 connections are allowed to complete their jobs and terminate naturally.

20        *Start* – A command verb instructing the SIP technology to put a server or  
21 cluster into an active state so that it accepts load on behalf of the cluster.

22        *Stop* – A command verb instructing the SIP technology to put a server or  
23 cluster into an inactive state so that it no longer accepts load on behalf of the  
24 cluster. The accompanying state change does not guarantee completion of open  
25 connections.

1        *Watermark* – Percentage of the cluster that must remain active even if the  
2 servers are failing application-layer testing. For example, to maintain two servers  
3 in a three-server cluster the watermark should any value between 34% and 66%.  
4 The exemplary monitor/controller does not add a failing server in order to meet  
5 the cluster watermark, because the watermark is only used to determine if a failing  
6 server can be removed. Furthermore, suspended servers count against the number  
7 of servers that can be removed.

### 8 9        **Exemplary Monitor/Controller**

10        Fig. 1 illustrates an example of a load-balancing cluster 100 and clients  
11 (such as 132-138) accessing such cluster via the Internet 120 (or any network for  
12 that matter). The exemplary monitor/controller may be implemented in this  
13 environment. More specifically, the exemplary monitor/controller may be  
14 implemented at an apparent client workstation (such as 132-138). Since the  
15 exemplary monitor/controller is remote from the cluster that it monitors, it may be  
16 any client external from the cluster. It may be located physically adjacent the  
17 cluster or it may be located on the other side of the globe.

18        For illustration purposes, client 132 is designated “monitor workstation”  
19 132 and it implements the exemplary monitor/controller. Since an implementation  
20 of the exemplary monitor/controller is modular, the functions (i.e., components) of  
21 the exemplary monitor/controller may be spread across multiple computers.  
22 Therefore, the monitor workstation 132 may represent multiple computers that are  
23 communicatively linked. In addition, the monitor workstation may monitor and/or  
24 control any node 112a-f of the cluster 100. Furthermore, the monitor workstation  
25 may monitor and/or control multiple clusters.

## Exemplary Monitor/Controller Architecture

Fig. 2 illustrates an architecture of the exemplary monitor/controller. Primarily, it includes components 210-244 within the monitor workstation 132. Fig. 2 also includes exemplary clusters 270 and 280 to illustrate the focus of the monitoring and control of the exemplary monitor/controller.

The exemplary monitor/controller is built as a set of cooperating components, resulting in an architecture that is open and flexible. Fig. 2 illustrates the high-level components of the system as well as the way they interact. The components are an admin station 242, reporter station 244, database 240, and control system 205.

The admin station 242 is a user-interface component that operators use to interact with the rest of the system. In the exemplary monitor/controller, all communication from the admin station 242 proceeds through the database 240. There is no direct communication between the admin station 242 and the control system 205.

In the exemplary monitor/controller, the reporter 244 is a set of account-protected Web pages from which current status and availability information can be obtained. The Web pages of the reporter can be accessed directly through a browser or through a redirect in the admin station 242.

1 The database 240 is the central storage unit of the exemplary  
2 monitor/controller. An example of such a database is an SQL database like that  
3 provided by Microsoft® SQL Server®. The database serves at least three  
4 purposes:

- 5 • *Configuration Storage* – all information needed to monitor clusters and  
6 their nodes is stored here.
- 7 • *State information repository* – as state changes occur (e.g., a node that fails,  
8 then passes testing), information is stored in the database to record what  
9 happened and when. This provides the current status of the servers and is  
10 also used for availability analysis.
- 11 • *Communication link* – one can control the operation and behavior of the  
12 control system 205 by issuing actions from the admin station 242. Actions  
13 are communicated to the control system 205 through the database which  
14 give it a specific task to perform (e.g., shut itself down or reload its  
15 configuration information).

## 16 Control System

17  
18 As shown in Fig. 2, the control system 205 includes several components:  
19 Controller 210; monitor framework 222; app-monitors (such as 226 and 228); SIP  
20 API interfaces 232, and cluster-controls (such as 236 and 238).

21 As its name implies, the controller 210 is the central component of the  
22 control system 205. Examples of some the tasks it performs includes coordinating  
23 information about servers from the application layer and their state in a SIP  
24 cluster; determining whether a node should participate in cluster activity; and  
25 retrieving information from and writing information to the database.



1 The monitor framework 222 is a module for managing the app-monitors  
2 and other monitoring components. It assists the controller 210 in its duties. It is  
3 an insulator that hides the complexities of assessing service health.

4 The app-monitors (such as 226 and 228) plug into the monitor framework  
5 222—one app-monitor for each supported protocol. Examples of such protocols  
6 include HTTP, SMTP, OCBC SQL, etc. These app-monitors actually perform the  
7 monitoring of nodes of clusters. For example, as illustrated in Fig. 2, monitor-1  
8 226 monitors the nodes 282 of cluster-1 280 using the HTTP protocol. In another  
9 example, monitor-N 228 monitors the nodes 272 of cluster-N 270 using the SMTP  
10 protocol.

11 The SIP API interface 232 is a set of application programming interfaces  
12 (APIs) that are designed to facilitate cluster communication. It encapsulates all  
13 interactions with SIP technologies and is used to control and query for the status of  
14 cluster members.

15 The cluster-controls (such as 236 and 238) plug into the SIP APIs 232—  
16 one cluster-control for each supported clustering technology. These cluster-  
17 controllers actually perform the controlling and querying of nodes of clusters. For  
18 example, control-1 236 monitors cluster state and controls cluster participation of  
19 the nodes 282 in cluster-1 280. In addition, cluster-1 280 is monitored (both the  
20 cluster and the nodes 282) at the application-layer by monitor-1. In another  
21 example, control-N monitors cluster state and controls cluster participation of the  
22 nodes 272 in cluster-N 270. Monitor-N 228 monitors the application-layer state of  
23 cluster-N 270 and its nodes 272. With the framework 222 and SIP APIs 232, the  
24 controller 210 need not know about the protocols or tests involved to assess health.  
25 As new monitoring components come along, they snap into the control system

1 transparently. In addition, the controller 210 can make health determinations by  
2 combining the test results from more than one monitor.

3 The control system 205 is where the results of application-layer monitoring  
4 are reconciled with the current state of the nodes in the cluster. Control of cluster  
5 state is performed, removing servers failing application-layer monitoring and  
6 adding servers passing application-layer monitoring.

7 The lifetime of a control system is managed by a service of the operating  
8 system of the monitor workstation 132. The service retrieves the list of needed  
9 control systems from the database 240, then instantiates each system in its own  
10 process. The control system 205 then reads its profile from the database and  
11 begins monitoring.

12 A control system profile consists of all information needed for it to monitor  
13 and control a set of clusters. This includes cluster properties (e.g., IP address) and  
14 the tests that monitor application-layer health. Note that the service can host one or  
15 more control systems, a control system can monitor and control many clusters, and  
16 a cluster can have many tests performed on its servers.

### 17 **Methodological Implementation of the Exemplary Monitor/Controller**

18  
19 Fig. 3 shows a methodological implementation of the exemplary  
20 monitor/controller performed by the exocluster application-layer  
21 monitoring/controlling system 205 (or some portion thereof). This methodological  
22 implementation may be performed in software, hardware, or a combination  
23 thereof.

24 At 310 of Fig. 3, the exemplary monitor/controller is instructed to monitor a  
25 given cluster. At 312, it dynamically determines the membership of the given

1 cluster and the cluster state of each member of the cluster. It may do so via a  
2 “query” command.

3 At 314, it exocusterly monitors the members of the cluster from a client-  
4 perspective. In particular, it monitors the members of the cluster at the  
5 application-layer.

6 Such monitoring is achieved by testing the application layer. There are zero  
7 or more tests to assess application-layer health. A test may include the following  
8 properties:

- 9 • ID – A unique identifier for the test assigned by the database.
- 10 • Name – A friendly name to identify the test.
- 11 • Description – Friendly text describing the test.
- 12 • Test String – The information a monitor sends to test the server (  
13 e.g., /default.htm for a web server).
- 14 • Type – A code used by the monitor to tell it what to do with the test  
15 string (e.g., *Get Data* for performing an HTTP GET against a web  
16 server).
- 17 • Port # – The server-side port to receive the test string.
- 18 • Timeout – The number of milliseconds to wait for the test to  
19 complete before aborting and declaring the test failed.
- 20 • Retries – The number of additional consecutive, failed attempts the  
21 monitor will make before declaring an error.
- 22 • Interval – How often (milliseconds) the test should be performed.
- 23 • Monitor – A code to identify the type of monitor.
- 24 • Authentication Username and Password – Optional fields used if the  
25 test requires authentication.

- Authentication Type – the type of authentication to be used for the testing. Examples are anonymous (no authentication), clear text and NTLM.

In addition to testing for a response from a node as an indication of application-layer availability status (i.e., application-layer state), the exemplary monitor/controller may test in other ways.

One way is by simply examining a code returned by the node. This code is the result of a node's self-assessment of its availability. This code is typically tucked away in header information, but it may be in the body also.

Another way is to perform customized tests. In this way, the exemplary monitor/controller passes customized parameters to the node with the expectation of receiving known correct results (or a range of results) from the node. The node passes the test if the results returned are correct. For example, assume that a node hosts a search engine. It is known that given particular search terms that the node should return a given number (or specific type) of results.

At 320, the exemplary monitor/controller determines whether a member of the cluster is unavailable at the application-layer. If a node is overwhelmed, the exemplary monitor/controller sends a signal to stop that node at 322. Alternatively, it may signal that the overwhelmed node be placed into the “drain” state—which is a graceful way to “empty the check-out line”. It may also alert the operator that the node needs servicing.

At 330 of Fig. 3, the exemplary monitor/controller determines whether a presently inactive member of the cluster is now available at the application-layer. Such a member may be inactive because it was stopped or drained earlier when it

1 was overloaded. If an inactive node is otherwise ready for additional load, the  
2 exemplary monitor/controller sends a signal to start that node at 332.

3 At 340, the exemplary monitor/controller tracks statistics related to the  
4 present and historical availability and activity of the members of the cluster. It  
5 stores these statistics into a database (such as database 240). These statistics may  
6 be viewed by a reporter station (such as reporter 244).

7 At 342 of Fig. 3, the exemplary monitor/controller identifies  
8 activity/inactivity cycles. In such a cycle, a node is regularly swamped after being  
9 activated. Thus, this node is caught in a vicious cycle of activation, overload,  
10 inactivation, activation, overload, inactivation, and so forth. Using the recorded  
11 tracking statistics of recent history, the exemplary monitor/controller identifies this  
12 vicious cycle and attempts to stop it. It may wait an additional length of time  
13 before reactivating. It may take the member permanently offline and alert the  
14 operator that the node needs servicing. Step 332 (where an available node is  
15 started) may skip nodes that are identified as being part of a vicious cycle. At 344,  
16 the exemplary monitor/controller reports current statistics upon request. Likewise,  
17 it reports historical statistics upon request.

18 The process repeats as long as the members of one or more clusters are  
19 being monitored and controlled. Some steps may be performed concurrently.  
20

### 21 **More on the Exemplary Monitor/Controller**

22 Active Monitoring. Unlike conventional exocluster application-layer  
23 monitors (which are purely passive monitors), the exemplary monitor/controller  
24 controls the members of the cluster as well as monitors them. They can stop a  
25

1 problem node. Moreover, they can start an inactive node once its problems have  
2 been resolved.

3 Protocol Agnostic. Unlike conventional exocluster application-layer  
4 monitors, the exemplary monitor/controller is not limited to a specific protocol. It  
5 can monitor any protocols without being reprogrammed. Examples of such  
6 protocols include HTTP, Chat, SMTP, SQL, and a generic TCP port-connect  
7 monitor for protocols not currently supported.

8 Dynamic Adjustment to Cluster Membership. Unlike conventional  
9 exocluster application-layer monitors that monitor only a static set of members of  
10 a cluster, the exemplary monitor/controller can monitor and control the entire  
11 membership of a cluster as that membership dynamically changes. In other words,  
12 it discovers and begins monitoring members as they are added to the cluster and  
13 stops the monitoring as they are removed from the cluster.

14 Tracking Historical Availability and Activity States. Unlike conventional  
15 exocluster application-layer monitors, the exemplary monitor/controller tracks and  
16 stores historical data regarding the application-layer test results and changes in the  
17 cluster state of the cluster nodes. Analysis of this data can be made to determine  
18 application-layer availability (known as an availability analysis) and determine the  
19 significance of each failure type (known as a failure analysis).

20 This data can also be used to locate and highlight common errors. This  
21 may be generally called “trending.” These trends may be reported for manual or  
22 automated analysis. Analyzing trends in this way provides a mechanism to  
23 prevent similar errors occurring in similar cluster configurations.

## Other Additional Details

There is a “one to many” relationship between the controller 210 and an application-layer app-monitor (such as app-monitors 226 and 228). For an app-monitor instance, there is exactly one controller instance to which it is responsible. On the other hand, a controller can coordinate the information from more than one app-monitor in order to determine node health. There may be multiple controllers running at the same time.

With the exemplary monitor/controller, the control system 205 runs under a service of the operating system. The service will point to a database (such as database 240) from which it obtains a configuration file for a controller. The controller (such as controller 210) loads its configurations from the database on startup in order to initialize itself. Alongside other information are the types of app-monitors that are configured for the controller to do its job. It will use this information in order to initialize the app-monitors of the system. This configuration will provide the controller 210 with all of the information it needs to start up. Examples of configuration information are: the protocol to use for application-layer monitoring, how to assess health using the application-layer protocol, test interval, timeout period for testing, the IP addresses of the cluster, its clustering technology, the number of hosts expected for each cluster, whether they are to be controlled or just monitored, and customer/owner contact information by cluster .

On startup, the controller 210 requests an interface to a set of app-monitors via framework 222. Information in the database identifies the clusters to be monitored by the controller, which avoids the same cluster from being controlled

1 by more than one controller. The controller 210 can control/monitor an arbitrary  
2 number of clusters with arbitrary cluster size. Some clusters can be placed in  
3 “read-only” mode, in which the state is monitored, but not modified. In addition,  
4 non-Single IP technology servers can be monitored for availability measurements.

5 The controller 210 periodically checks the membership of the cluster and  
6 the cluster-state of the members. This is the common loop run through for the  
7 controller 210. The application-layer state is checked from the app-monitors for  
8 the cluster IP address and the node addresses. Any needed cluster-state changes  
9 are made to put in healthy servers and remove sick ones. The controller 210  
10 persists both cluster-state and application-layer state changes over time so that a  
11 person can look for trends indicative of problems not evident from a single  
12 snapshot of the status.

13 Information from poll results and cluster membership is stored for offline  
14 viewing, but in an efficient manner such that only changes in state are persisted.  
15 Server and cluster status (this includes monitor information about the servers,  
16 cluster state information, and cluster availability estimates) is logged to the  
17 database 240, typically as the data is produced.

18 The controller 210 learns cluster membership and the cluster-state of each  
19 server directly from the cluster. This discovery is part of the normal loop for the  
20 controller 210, in which the cluster is queried for all of the members and their  
21 cluster-state. On each pass (including startup), this step is used to determine the  
22 list of servers that the app-monitors should track. The controller 210 only updates  
23 this list to the app-monitors when the list changes. When a node is no longer part  
24 of the cluster, the monitoring continues for a configurable amount of time, to give  
25



1 the host a chance to return (maybe it is being rebooted). If it doesn't the host is  
2 removed from monitoring at the application-layer.

3 The controller 210 passes a list of IP addresses (including a VIP) for the  
4 app-monitor to track. The controller 210 then periodically requests the status of  
5 the monitored addresses in the form of state changes. The application-layer  
6 monitor returns only what has changed since the last time the controller polled for  
7 the status. An application-layer state change consists of information such as IP  
8 address tested, protocol independent label of the test performed, protocol-  
9 dependent and protocol-independent status codes, time stamp of the state change  
10 and protocol-dependent information returned as a string (e.g., a "HTTP 200" result  
11 description).

12 If a healthy host is in the stopped state, the controller 210 will start it, if the  
13 controller is configured to control the state of this cluster's hosts. If a sick server is  
14 in the converged state, the controller 210 may drain, wait, and then stop. However,  
15 if removing the server will drop the number of servers below the "watermark", the  
16 server is not removed. Watermarks are set through the configuration on a cluster-  
17 by-cluster basis as a fraction of the expected cluster size. The drain portion can be  
18 bypassed (e.g., if the wait time is set to zero in the configuration for the cluster).

19 If a server is in an admin "Suspended" state, the controller 210 monitors it,  
20 but does not control it.

21 The controller 210 detects when it is in a vicious cycle between activating  
22 and deactivating a node. State is maintained in the controller 210 for each cluster  
23 in order to catch problems visible only by tracking state changes over time. For  
24 example, if a server is oscillating in and out of service (or in and out of the  
25 cluster), the controller 210 discovers this as an error condition.

## **Exemplary Computing System and Environment**

Fig. 4 illustrates an example of a suitable computing environment 900 within which an exemplary monitor/controller, as described herein, may be implemented (either fully or partially). The computing environment 900 may be utilized in the computer and network architectures described herein.

The exemplary computing environment 900 is only one example of a computing environment and is not intended to suggest any limitation as to the scope of use or functionality of the computer and network architectures. Neither should the computing environment 900 be interpreted as having any dependency or requirement relating to any one or combination of components illustrated in the exemplary computing environment 900.

The exemplary monitor/controller may be implemented with numerous other general purpose or special purpose computing system environments or configurations. Examples of well known computing systems, environments, and/or configurations that may be suitable for use include, but are not limited to, personal computers, server computers, thin clients, thick clients, hand-held or laptop devices, multiprocessor systems, microprocessor-based systems, set top boxes, programmable consumer electronics, network PCs, minicomputers, mainframe computers, distributed computing environments that include any of the above systems or devices, and the like.

The exemplary monitor/controller may be described in the general context of computer-executable instructions, such as program modules, being executed by a computer. Generally, program modules include routines, programs, objects, components, data structures, etc. that perform particular tasks or implement

1 particular abstract data types. The exemplary monitor/controller may also be  
2 practiced in distributed computing environments where tasks are performed by  
3 remote processing devices that are linked through a communications network. In  
4 a distributed computing environment, program modules may be located in both  
5 local and remote computer storage media including memory storage devices.

6 The computing environment 900 includes a general-purpose computing  
7 device in the form of a computer 902. The components of computer 902 can  
8 include, by are not limited to, one or more processors or processing units 904, a  
9 system memory 906, and a system bus 908 that couples various system  
10 components including the processor 904 to the system memory 906.

11 The system bus 908 represents one or more of any of several types of bus  
12 structures, including a memory bus or memory controller, a peripheral bus, an  
13 accelerated graphics port, and a processor or local bus using any of a variety of  
14 bus architectures. By way of example, such architectures can include an Industry  
15 Standard Architecture (ISA) bus, a Micro Channel Architecture (MCA) bus, an  
16 Enhanced ISA (EISA) bus, a Video Electronics Standards Association (VESA)  
17 local bus, and a Peripheral Component Interconnects (PCI) bus also known as a  
18 Mezzanine bus.

19 Computer 902 typically includes a variety of computer readable media.  
20 Such media can be any available media that is accessible by computer 902 and  
21 includes both volatile and non-volatile media, removable and non-removable  
22 media.

23 The system memory 906 includes computer readable media in the form of  
24 volatile memory, such as random access memory (RAM) 910, and/or non-volatile  
25 memory, such as read only memory (ROM) 912. A basic input/output system

1 (BIOS) 914, containing the basic routines that help to transfer information  
2 between elements within computer 902, such as during start-up, is stored in ROM  
3 912. RAM 910 typically contains data and/or program modules that are  
4 immediately accessible to and/or presently operated on by the processing unit 904.

5 Computer 902 may also include other removable/non-removable,  
6 volatile/non-volatile computer storage media. By way of example, Fig. 4  
7 illustrates a hard disk drive 916 for reading from and writing to a non-removable,  
8 non-volatile magnetic media (not shown), a magnetic disk drive 918 for reading  
9 from and writing to a removable, non-volatile magnetic disk 920 (e.g., a “floppy  
10 disk”), and an optical disk drive 922 for reading from and/or writing to a  
11 removable, non-volatile optical disk 924 such as a CD-ROM, DVD-ROM, or other  
12 optical media. The hard disk drive 916, magnetic disk drive 918, and optical disk  
13 drive 922 are each connected to the system bus 908 by one or more data media  
14 interfaces 926. Alternatively, the hard disk drive 916, magnetic disk drive 918,  
15 and optical disk drive 922 can be connected to the system bus 908 by one or more  
16 interfaces (not shown).

17 The disk drives and their associated computer-readable media provide non-  
18 volatile storage of computer readable instructions, data structures, program  
19 modules, and other data for computer 902. Although the example illustrates a  
20 hard disk 916, a removable magnetic disk 920, and a removable optical disk 924,  
21 it is to be appreciated that other types of computer readable media which can store  
22 data that is accessible by a computer, such as magnetic cassettes or other magnetic  
23 storage devices, flash memory cards, CD-ROM, digital versatile disks (DVD) or  
24 other optical storage, random access memories (RAM), read only memories  
25 (ROM), electrically erasable programmable read-only memory (EEPROM), and

1 the like, can also be utilized to implement the exemplary computing system and  
2 environment.

3 Any number of program modules can be stored on the hard disk 916,  
4 magnetic disk 920, optical disk 924, ROM 912, and/or RAM 910, including by  
5 way of example, an operating system 926, one or more application programs 928,  
6 other program modules 930, and program data 932. Each of such operating  
7 system 926, one or more application programs 928, other program modules 930,  
8 and program data 932 (or some combination thereof) may include an embodiment  
9 of a dynamic cluster-membership determiner, an exocluster application-layer  
10 monitor, an exocluster controller, an overload-identifier, a state-determiner, a  
11 database, an app-monitor, cluster-control, and a central controller.

12 A user can enter commands and information into computer 902 via input  
13 devices such as a keyboard 934 and a pointing device 936 (e.g., a "mouse").  
14 Other input devices 938 (not shown specifically) may include a microphone,  
15 joystick, game pad, satellite dish, serial port, scanner, and/or the like. These and  
16 other input devices are connected to the processing unit 904 via input/output  
17 interfaces 940 that are coupled to the system bus 908, but may be connected by  
18 other interface and bus structures, such as a parallel port, game port, or a universal  
19 serial bus (USB).

20 A monitor 942 or other type of display device can also be connected to the  
21 system bus 908 via an interface, such as a video adapter 944. In addition to the  
22 monitor 942, other output peripheral devices can include components such as  
23 speakers (not shown) and a printer 946 which can be connected to computer 902  
24 via the input/output interfaces 940.  
25

1 Computer 902 can operate in a networked environment using logical  
2 connections to one or more remote computers, such as a remote computing device  
3 948. By way of example, the remote computing device 948 can be a personal  
4 computer, portable computer, a server, a router, a network computer, a peer device  
5 or other common network node, and the like. The remote computing device 948 is  
6 illustrated as a portable computer that can include many or all of the elements and  
7 features described herein relative to computer 902.

8 Logical connections between computer 902 and the remote computer 948  
9 are depicted as a local area network (LAN) 950 and a general wide area network  
10 (WAN) 952. Such networking environments are commonplace in offices,  
11 enterprise-wide computer networks, intranets, and the Internet.

12 When implemented in a LAN networking environment, the computer 902 is  
13 connected to a local network 950 via a network interface or adapter 954. When  
14 implemented in a WAN networking environment, the computer 902 typically  
15 includes a modem 956 or other means for establishing communications over the  
16 wide network 952. The modem 956, which can be internal or external to computer  
17 902, can be connected to the system bus 908 via the input/output interfaces 940 or  
18 other appropriate mechanisms. It is to be appreciated that the illustrated network  
19 connections are exemplary and that other means of establishing communication  
20 link(s) between the computers 902 and 948 can be employed.

21 In a networked environment, such as that illustrated with computing  
22 environment 900, program modules depicted relative to the computer 902, or  
23 portions thereof, may be stored in a remote memory storage device. By way of  
24 example, remote application programs 958 reside on a memory device of remote  
25 computer 948. For purposes of illustration, application programs and other

1 executable program components such as the operating system are illustrated herein  
2 as discrete blocks, although it is recognized that such programs and components  
3 reside at various times in different storage components of the computing device  
4 902, and are executed by the data processor(s) of the computer.

### 5 **Computer-Executable Instructions**

7 An implementation of an exemplary monitor/controller may be described in  
8 the general context of computer-executable instructions, such as program modules,  
9 executed by one or more computers or other devices. Generally, program modules  
10 include routines, programs, objects, components, data structures, etc. that perform  
11 particular tasks or implement particular abstract data types. Typically, the  
12 functionality of the program modules may be combined or distributed as desired in  
13 various embodiments.

### 14 **Exemplary Operating Environment**

16 Fig. 4 illustrates an example of a suitable operating environment 900 in  
17 which an exemplary monitor/controller may be implemented. Specifically, the  
18 exemplary monitor/controller(s) described herein may be implemented (wholly or  
19 in part) by any program modules 928-930 and/or operating system 928 in Fig. 4 or  
20 a portion thereof.

21 The operating environment is only an example of a suitable operating  
22 environment and is not intended to suggest any limitation as to the scope or use of  
23 functionality of the exemplary monitor/controller(s) described herein. Other well  
24 known computing systems, environments, and/or configurations that are suitable  
25 for use include, but are not limited to, personal computers (PCs), server

1 computers, hand-held or laptop devices, multiprocessor systems, microprocessor-  
2 based systems, programmable consumer electronics, wireless phones and  
3 equipments, general- and special-purpose appliances, application-specific  
4 integrated circuits (ASICs), network PCs, minicomputers, mainframe computers,  
5 distributed computing environments that include any of the above systems or  
6 devices, and the like.

### 7 **Computer Readable Media**

8  
9 An implementation of an exemplary monitor/controller may be stored on or  
10 transmitted across some form of computer readable media. Computer readable  
11 media can be any available media that can be accessed by a computer. By way of  
12 example, and not limitation, computer readable media may comprise “computer  
13 storage media” and “communications media.”

14 “Computer storage media” include volatile and non-volatile, removable and  
15 non-removable media implemented in any method or technology for storage of  
16 information such as computer readable instructions, data structures, program  
17 modules, or other data. Computer storage media includes, but is not limited to,  
18 RAM, ROM, EEPROM, flash memory or other memory technology, CD-ROM,  
19 digital versatile disks (DVD) or other optical storage, magnetic cassettes, magnetic  
20 tape, magnetic disk storage or other magnetic storage devices, or any other  
21 medium which can be used to store the desired information and which can be  
22 accessed by a computer.

23 “Communication media” typically embodies computer readable  
24 instructions, data structures, program modules, or other data in a modulated data  
25



1 signal, such as carrier wave or other transport mechanism. Communication media  
2 also includes any information delivery media.

3 The term “modulated data signal” means a signal that has one or more of its  
4 characteristics set or changed in such a manner as to encode information in the  
5 signal. By way of example, and not limitation, communication media includes  
6 wired media such as a wired network or direct-wired connection, and wireless  
7 media such as acoustic, RF, infrared, and other wireless media. Combinations of  
8 any of the above are also included within the scope of computer readable media.

### 9 **Conclusion**

10  
11 Although the invention has been described in language specific to structural  
12 features and/or methodological steps, it is to be understood that the invention  
13 defined in the appended claims is not necessarily limited to the specific features or  
14 steps described. Rather, the specific features and steps are disclosed as preferred  
15 forms of implementing the claimed invention.  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25